Lecture No. 5
Probability Distribution

January 20, 2003

- Random variable: It is a function defined on the sample space.
  Example: Tossing a coin yield a sample space $S = \{H, T\}$.
  Let X be a random variable defined on S as $X : S \to \{0, 1\}$ i.e.,
  $X(c) = 0$ when $c = H$
  $X(c) = 1$ when $c = T$

- The random variable is a function that can take on either a finite number of values, each with an associated probability, or an infinite number of values whose probabilities are summarized by a density function.

- E.g., random variable defined on the outcomes in rolling of a die is discrete.

- E.g., random variable defined on the outcomes in measurement of the height of students in a class is continuous.

- Probability Distribution: A discrete probability distribution of a random variable X is the set of values that this random variable can take, together with their associated probabilities.
  Example: Tossing a fair coin twice, let random variable X be the number of heads that eventually turns up. The sample space is

$$S = \{HH, HT, TH, TT\}$$

  The possible values of X are 0, 1, 2.

$$\begin{aligned} X(c) &= 0 \quad \text{when} \quad c = TT \\ &= \quad 1 \text{ when} \quad c = HT, TH \\ &= \quad 2 \text{ when} \quad c = HH \end{aligned}$$

  The probability distribution of X is

X	0	1	2
$p(X = x)$	.25	.50	.25

- Some Important Distributions used in Bioinformatics:

Discrete	Binomial(n,p) Poisson(λ) Multinomial($n_1, \ldots, n_r, p_1, \ldots, p_r$)
Continuous	Normal(μ, σ) Extreme value

- Binomial(n,p): Toss a coin n times, what is the probability of having x heads? Assuming each tossing is independent of others and the probability of getting a head in a particular tossing is p. Also

1). total number of possibilities i.e., sample space.

2). the probability of a particular outcome with x heads and $(n - x)$ tails. The ordering does not matter.

$$p(x \text{ heads}) = C_x^n p^x (1-p)^{n-x} \quad x = 0, 1, 2, \ldots, n$$

- Example: Probability of finding one adenine (A) in 10 nucleotides, where probability for an individual nucleotide is 0.25
  i.e.,
  $p(A) = 0.25$ and $p(\text{not}A) = 0.75$.
  $x = 0, 1, 2, \ldots, 10$

$$\begin{aligned} p(x = 1) &= C_1^{10}(0.25)^1(1 - .25)^{10-1} \\ &= \frac{10!}{9!1!}(.25)^1(.75)^9 \\ &= 10(.25)(.75)^9 \\ &= 0.2 \end{aligned}$$

- Examples of Binomial Distribution:

1). Promoter region, not promoter region.

2). coding region or not coding region.

3). Alpha helix or not alpha helix.

- Poisson Distribution: Another distribution based on Poisson process is a probabilistic mechanism giving rise to the occurrence of events in a specific time interval (or region of space)

1). the probability of occurrence in an infinitesimally small area (or time interval) is very small

2). the probability that more than one event occurs in the small area/time is negligible

3). events over disjoint regions are independent

4). The constant μ is the rate parameter for Poisson distributed random variable.

- For a random variable $X \sim$ Poisson(μ)

  $$p(X = m) = \frac{\mu^m \exp(-\mu)}{m!} \quad m = 0, 1, 2, ...$$

  where e is the base of the natural logarithms. Its value is 2.718282. The poisson distribution is equivalent to the binomial in limit of an infinite number of trials (and a proportionately lower success rate per trial).

- Example: Number of restriction sites for a given enzyme in a large DNA molecule of random sequence.
  For a sufficiently large molecule, the potential number of sites is very large (essentially one per basepair).
  So poisson provides an accurate representation of the distribution of the number of sites in random sequence.

Example: In a 10240 nucleotide molecule, the expected number of sites for EcoRI is

$$\begin{aligned} \mu &= \frac{10240}{4^6} \\ &= 2.5 \end{aligned}$$

Thus the probability of no of EcoRI sites is

$$\begin{aligned} p(m=0) &= \frac{(2.5)^0 \exp(-2.5)}{0!} \\ &= \exp(-2.5) \\ &= 0.082 \end{aligned}$$

where μ is the mean and the standard deviation is $\sqrt{\mu}$.

- Multinomial Distribution: Extension of binomial distribution for many nucleotides. Probability of finding a certain pattern of bases.

- Examples:

1). Is a certain region coding or not coding.

2). Is it an intron or exon region.

3). What is the probability that a promoter of length "X" could be found within a certain length of a sequence.

- Multinomial answers the questions where there are more than two outcomes
  Alpha helix, beta strand, turn and other structures.
  Exon, intron, promoter and other regions.

- Multinomial Distribution:

  $$p(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

  provided that
  $n = \sum_{i=1}^{k} n_i$
  $\sum_{i=1}^{k} p_i = 1$

- Intron: DNA sequence that interrupts the protein coding sequence of a gene. An intron is transcribed into RNA but is cut out of the message before it is transcribed into protein.

- Exon: The protein coding DNA sequence of a gene.

- Promoter: A DNA site to which RNA polymerase binds and initiated transcription.

- Example: Suppose that all of the amino acids are equally likely (which is of course never true, but it makes calculation easier). Suppose we observed 5 amino acids selected independently at random. Then what is the probability of seeing 2 leuines, 2 isoleusine and 1 valine?

$$\begin{aligned} p(2,2,1) &= \frac{5!}{2!2!1!}(1/20)^2(1/20)^2(1/20)^1 \\ &= \frac{120}{4}(1/20)^5 \\ &= \frac{30}{20^5} \end{aligned}$$

- Continuous Distribution: Possible outcomes take on a continuum of possible values. Temperature at a certain time and place.
  Height of a randomly chosen person.
  Fluorescence intensity at a spot on a microarray etc.

1. For a continuous random variable, any particular value has probability 0 of occurring.

2. The density as the height of the histogram for the random variable (called the density curve).

3. The total area under the density curve = 1.

- A probability density function (PDF) $f(x)$ such that the probability between a range a to $b\,(a \leq b)$ is the integral of $f(x)$ over the interval from a to b.

$$p(a \leq x \leq b) = \int_a^b f(x)dx$$

- The (cumulative) distribution function (cdf) for (any) RV X is $F(x) = P(X \leq x)$.

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(y)dy \\ f(x) &= \frac{d(F(x)}{dx} \end{aligned}$$

The cdf satisfies:

1. F(x) is nondecreasing for all x.

2. $F(-\infty) = 0$.

3. $F(\infty) = 1$.

- Normal Distribution: Normal distribution is a smooth bell shaped curve and is symmetric about mean μ. Its shape (spread and dispersion) is characterized by the standard deviation σ. The distribution is completely determined by these two parameters μ and σ.

- Let x be a random variable (e.g., birth weight) having normal distribution with mean μ and standard deviation σ, $N(\mu, \sigma^2)$, then the probability density function is:

  $$f(x) = 1/\sqrt{2\pi\sigma^2}\,\exp\{\frac{-(x-\mu)^2}{2\sigma^2}\}\ -\infty \leq x \leq \infty$$

  where $-\infty \leq \mu \leq \infty$, $\sigma > 0$.

- The probability $p(X < x)$ is the area under the curve to the left of the value x.

- Some facts:

1). $p(\mu - \sigma < x < \mu + \sigma) = 68\%$

2). $p(\mu - 2\sigma < x < \mu + 2\sigma) = 95\%$

3). The normal distribution with 0 mean and standard deviation one (i.e., $N(0, 1)$) is called standard normal distribution, the associated random variable denoted by Z.

- The standard normal table is provided by most of the text books. For example from table:

$$p(Z < 1.96) = 0.025$$

- Probability for any other $N(\mu, \sigma^2)$ distribution can be obtained from the normal table using following transformation:

$$Z = \frac{x - \mu}{\sigma}$$

- Example: Systolic blood pressure (BP) is normally distributed with mean 120 and standard deviation 10. Find the probability that a person has BP greater than 130.

  Let X be a random variable distributed as $N(120, 100)$.

$$\begin{aligned} p(X > 130) &= p(\frac{X - \mu}{\sigma} > \frac{130 - 120}{10}) \\ &= p(Z > 1) \\ &= 0.1587 \end{aligned}$$

- Importance of the normal distribution in statistics

1. Convenient mathematical properties.

2. Variations in a number of physical experiments are often approximately normally distributed.

3. Central Limit Theorem (CLT), which says that if a sufficiently large random sample is taken from some distribution, then even though this distribution is not itself approximately normal, the distribution of the sample SUM or AVERAGE will be approximately normal.

- QQ-Plot Quantile-quantile plot

1. Used to assess whether a sample follows a particular (e.g. normal) distribution (or to compare two samples).

2 A method for looking for outliers when data are mostly normal.

3 Typical deviations from straight line patterns. Outliers Curvature at both ends (long or short tails).